Arabian Journal for Science and Engineering
https://doi.org/10.1007/s13369-019-03920-9

RESEARCH ARTICLE -COMPUTER ENGINEERING AND COMPUTER SCIENCE

On Term Frequency Factor in Supervised Term Weighting Schemes for Text Classification

Turgut Dogan[1] · **Alper Kursat Uysal**[1]

Received: 11 September 2018 / Accepted: 13 May 2019

Abstract
The performance of text classification can be affected by the choice of appropriate term weighting scheme as well as other parameters. The terminology supervised term weighting scheme has become popular in recent years, as it may provide discriminative representation in vector space for text documents belonging to different classes. A term weighting scheme generally consists of three factors, namely term frequency factor, collection frequency factor, and length normalization factor. The researchers mostly have been focused on developing new collection frequency factors in term weighting studies. However, the term frequency factor has an important role, especially in supervised term weighting. In this study, we extensively analyzed the effects of using different term frequency factors on seven supervised term weighting schemes. While six of these supervised term weighting schemes were applied in the previous studies in the literature, we derived one of them from an existing feature selection method and it was not used as a weighting method before. This analysis is performed using SVM and Roccio classifiers on two widely known benchmark datasets with different characteristics. Experimental results showed that modification of term frequency factor in supervised term weighting schemes increased the performance of almost all weighting schemes. Also, term weighting schemes using square root function-based term frequency factor (SQRT_TF) are more successful than the ones using term frequency (TF) and logarithmic function-based term frequency (LOG_TF) factors. TF term frequency factor seems as the least effective one among three different term frequency factors according to the experimental results and statistical analysis.

Keywords Text classification · Term weighting · Term frequency factor · Feature selection

1 Introduction

Depending on the rapid development of information technology and the Internet, the amount of electronic data containing text documents has risen dramatically. This rise made effectively storing, organizing, and retrieving these text documents critical. Text classification is an important way for managing and organizing these data in addition to obtaining helpful information. In other words, text classification can be defined as the function of assigning textual documents to predefined categories. There exist various solution methods to improve the success ratio of text classification such as preprocessing [1], feature selection [2–7], feature (term) weighting [8] besides employing new classification approaches [9]. Term weighting is a research field that attracts attention of researchers' working on text classification in recent years. Text documents are generally represented by feature vectors in text classification, and these vectors consist of weights indicating the contribution of each term. Since each weight represents the importance of the corresponding term in the document, term weighting is a crucial step to increase the success of text classification. For this purpose, some term weighting methods (approaches or schemes) have been proposed in the text classification literature.

The essential and traditional scheme proposed for term weighting is TF-IDF (term frequency-inverse document frequency). Its effectiveness on the weighting process is proven in information retrieval studies [10]. However, since the class information of the documents is not used for the weighting process, it is not very well suited for text classification such as information retrieval. In the early 2000s, the idea of

✉ Turgut Dogan
turgutdogan@eskisehir.edu.tr

Alper Kursat Uysal
akuysal@eskisehir.edu.tr

[1] Department of Computer Engineering, Eskişehir Technical University, Eskisehir, Turkey

Published online: 21 May 2019

using class information obtained from training documents in term weighting process is proposed [11]. Thus, the concept of supervised term weighting scheme (STW) has emerged. Debole and Sebastiani employed three well-known feature selection methods, namely information gain, gain ratio, and Chi-square for term weighting by displacing the IDF factor in TF-IDF. They showed that the proposed supervised term weighting schemes outperformed traditional TF-IDF scheme for text classification. After publication of these promising results obtained using these STW schemes, more researchers have been focused on STW schemes. Deng et al. [8] compared the performance of four weighting methods on Reuters-21578 benchmark dataset with support vector machines (SVMs) classifiers. They stated that term frequency–Chi square (TF-CHI2) term weighting is the most effective one among these four term weighting methods. Lertnattee and Theeramunkong [12] showed that when innumerable most-class terms and few-class terms are acquired, combination of logarithmic function of inverse class frequency and term frequency is the most efficient. In this study, they called a term occurring in only few classes as few-class terms and a term appearing in many classes as most-class terms. Therefore, they used three functions of inverse class frequency named as square root (ICFSqrt), logarithmic (ICFLog), and linear (ICFLinear) to combine with TF and TF-IDF term weightings on the centroid classifiers. They performed their experiments using unigram and bigram models on the centroid classifiers. Lan et al. [13] proposed a new term weighting scheme, namely term frequency relevance frequency (tf.rf), which they implied will improve the terms' discriminating capability for text classification task. They emphasized that since it supplies more contribution while selecting the positive samples from the negative samples, the term which has high frequency and is in the positive category is more important than the one which has high frequency and is in the negative category. They used unsupervised and supervised schemes for comparing tf.rf scheme on the benchmark datasets with SVM and k-nearest neighbor (kNN) classifier. They concluded that tf.rf had consistently better performance in comparison with the other term weighting methods. Liu et al. [14] offered a new scheme which uses two critical information related to ratios, namely relevance indicators to deal with class imbalance problem. In this scheme, they calculate weights using intra-class distributions of terms apart from inter-class distributions. They showed that their proposed scheme provides a better representation for minor categories, while the success for major categories is not negatively affected. Altınçay and Erenel [15] presented a comprehensive analysis of six widely utilized term weighting schemes for text classification. They showed the weighting behaviors of schemes by analyzing the relation between the existence probabilities of terms which receive equal weights, and they clarified similarities and differences of schemes. Deisy et al. [16] proposed a novel term weighting scheme named as modified inverse document frequency (MIDF) to improve the success of text categorization. They tried to overcome the problem that the success of weighting scheme is significantly affected when the number of classes increased. The weighting scheme they introduced focuses on document frequency and term frequency, but it does not use the information about the number of documents in whole collection. They showed that the overall performance of MIDF scheme with SVM classifier is better than TF-IDF and weighted inverse document frequency (WIDF) schemes. Wei et al. [17] proposed a new approach for term weighting using semantic similarity. They stated that most researchers do not take into account the semantic relation between categories and terms on the weighting process. The weighting process of their method consists of three stages. In the first stage, they modeled the categories with two feature vectors. Then, they calculated the semantic distance between feature terms and category core terms with the help of the semantic database, namely WordNet in the second stage. Finally, they used semantic distance instead of global weight factor of conventional weighting schemes to acquire the weight of every term. They concluded that the STW schemes used with proposed approaches generally outperform other equivalents which do not use semantic similarity. Similarly, Luo et al. [18] declared that the weight of a specific term is associated with its semantic similarity to a category and they proposed a new semantic term weighting scheme for text categorization. They stated that their proposed approaches considerably outperform TF-IDF if the amount of training data is small or the content of documents is focused on well-defined categories. Ren and Sohrab [19] introduced two term weighting schemes consisting of the information of both document and class indexing to provide positive discrimination on infrequent and frequent terms in text classification. By multiplying the inverse class frequency (ICF) factor by TF-IDF, they created the first scheme called as TF-IDF-ICF. The second scheme, namely TF-IDF-ICS&F method, is generated by revising the ICF function and implementing inverse class space density frequency (ICS&F). They showed that the second proposed scheme obtains better results than other well-known term weighting schemes. Emmanuel et al. [20] proposed another term weighting approach, namely positive impact factor (PIF), which is a variation of traditional supervised term weighting methods. They were inspired by the assumption that 'Positive impact of a feature to a category can be utilized to calculate its negative impact for other categories' on the developing process of this weighting scheme. They showed that their proposed scheme improved accuracy and reduced computational cost significantly compared to existing term weighting methods such as Binary, TF, TF-IDF, TF-RF, and TF-CHI2.

Badawi and Altınçay [21] introduced a novel framework based upon working the joint occurrence statistics of pairs of terms for term set selection and weighting. They focused the idea that the existence of one term but not the other may also transfer precious information for discrimination. They weighted the selected term set according to this idea, and they concluded that the proposed idea can be effectively used for term set selection and weighting. Ke [22] presented information-theoretic term weighting schemes for document classification and clustering. Ke derived two basic quantities, namely LI binary (LIB) and LI frequency (LIF), for document representation. Ke used these quantities for weighting as individually and by combining them. Ke demonstrated that the LIB $\times$ LIF scheme is superior to TF-IDF. Deng et al. [23] proposed a novel supervised scheme for sentiment analysis. In the proposed scheme, the terms were weighted according to two factors: ITD and ITS. ITD shows significance of a term in a document (ITD), and ITS demonstrated significance of a term for expressing sentiment (ITS). They showed that their proposed scheme can make full use of the available labeling information to assign proper weights to terms compared with the previous unsupervised term weighting schemes. They also showed that the term weighting schemes focused upon supervised learning are more important than those schemes originated from information retrieval without taking into account the correlation between terms and sentiment polarity. In another study for sentiment classification, Fattah proposed three new term weighting schemes that benefit from class space density based upon the class distribution in all document sets as well as the class documents set [24]. Fattah compared these novel term weighting schemes with six traditional and state-of-art weighting schemes on the Gaussian mixture model (GMM), SVM, and probabilistic neural network (PNN) classifier. Fattah concluded that the success of some of the proposed schemes is better than other traditional and the latest weighting schemes for sentiment classification. Escalante et al. [25] proposed a genetic program which aims at learning more effective term weighting schemes. They reported that term weighting schemes constructed with this genetic program outperformed traditional schemes and other term weighting schemes proposed in recent works. They emphasized that if genetic programming is used usefully for term weighting, the performance of classification can be increased effectively. In the novel term weighting study, Ko calculated the odds of positive and negative class probabilities by using class information to find weights of terms [26]. They stated that the results of experiments where proposed scheme (TF-TRR) was compared with TF-RF, TF-IDF, and two different versions of TF-IDF proved the superiority of TF-TRR. Chen et al. [27] proposed a novel statistical model, namely inverse gravity moment (IGM), to characterize the inter-class distribution for term weighting. They proposed TF-IGM and square root TF-IGM (RTF) term weighting schemes relying on IGM model. They stated that these schemes aim to absolutely measure the class-specific discrimination ability of a term. The results on SVM and kNN classifiers showed that proposed schemes outperformed other traditional, supervised, and unsupervised schemes in terms of performance measures such as micro-F_1 and macro-F_1. Haddoud et al. [28] collected innumerable metric functions which were proposed for other data mining problems to measure the correlation between two events. Their aim was to measure the performances of these metrics for term weighting. So, they proposed 80 metrics for the term weighting problem which were never used in the term weighting literature. Since optimal results cannot be achieved from any metric according to whether the label distribution is balanced or not, they proposed a classifier combination method consisting of different metrics. Kim et al. [29] proposed novel term weighting schemes inspired from naïve Bayes and the multinomial term model. They tested their weighting approaches on four datasets and verified that the proposed methods showed significantly better performances over existing term weighting methods. Sabbah et al. [30] introduced four modified-frequency-based term weighting schemes derived from the standard TF, IDF, and TF-IDF schemes. The idea behind these schemes was to account for missing terms from documents while terms were weighted. They used TF, DF, TF-IDF Glasgow and entropy weighting schemes to convert proposed schemes on the Reuters-21578, 20 Newsgroups and WebKB datasets. They showed that proposed schemes achieved the maximum performance on the SVM classifier and outperformed weighting approaches such as TF, TF-IDF, and Entropy. In the recent years, some hypotheses have been verified on the assumptions that the occurrence of a subset of terms can be also given information about the class memberships for term weighting. Badawi and Altınçay proposed a term set weighting strategy which is focused on the cardinality statistics and this hypothesis [31]. They assumed that the occurrence of a subset of the members may also be discriminative, especially in the case of individually discriminative members. They used two existing collection frequencies to adopt term set weighting by their BOW-based representation method. They showed that BOW-based representation can be successfully enriched utilizing term sets that are weighted using the proposed scheme. Over-weighting, under-weighting, and imbalanced datasets are other challenge topics for term weighting. The researchers working on term weighting have also concentrated on these topics as well as proposing new term weighting approaches. For instance, unsuitable handling of singular terms and huge ratios between term weights are referred to as over-weighting. Wu et al. [32] implied that assigning large weights to imbalanced terms and controlling the balance between under-weighting and over-weighting can improve the success of supervised term weighting scheme. They focused

especially on over-weighting and presented three regularization methods named add-one smoothing, sublinear scaling, and bias term to avoid over-weighting. They used add-one smoothing to deal with improper handling of singular terms, while they benefited from sublinear scaling and bias term to narrow the ratios between term weights. They have also proposed a new supervised term weighting scheme, namely regularized entropy (re) employing entropy to measure term distribution and introducing the bias term to control under-weighting and over-weighting. They showed that their regularization methods increased significantly the success of supervised term weighting, and their proposed methods outperformed other existing approaches. Alsmadi et al. [33] offered a novel term weighting scheme, namely SW, for short text categorization. They used two datasets including various number of tweets. They evaluated their methodology with four learning algorithms, and tenfold cross-validation is carried out. The evaluation results indicated that their proposed scheme majorly outperformed other approaches in terms of accuracy. They stated that SW overcomes challenges of limitation, sparsity, and noise in short texts more than the other weighting methods.

The studies about term weighting which was lately carried out can be summarized as follows: Rao et al. [34] proposed a term weighting scheme for first story detection. Feng et al. [35] presented relevance (lrp) term weighting scheme based on a probabilistic model. Matsuo et al. [36] carried out a study consisting a two-stage framework which assigns weights to terms according to semantic relations for clinical text documents. Li et al. [37] proposed a combination of entropy weighting (CEW) with two existing schemes for effective topic modeling. Santhanakumar et al. [38] proposed co-term frequency term weighting method based on weighting multi-terms that jointly occur in all documents.

As mentioned in the previous paragraphs, the researchers mostly have been focused on developing new collection factors in order to propose new term weighting schemes. However, the term frequency factor in term weighting scheme has a significant role in supervised term weighting. Different term frequency factors are already applied in some studies, but their positive or negative impact on the corresponding methods is not investigated in detail [27, 39–41]. In this study, the impact of term frequency factor on the performance of supervised term weighting methods has been investigated. For this purpose, three term frequency factors which were used in different studies in the literature have been used in the weighting process. In order to verify the effect of term frequency factor, experiments were realized using seven supervised term weighting scheme and two classifiers on two benchmark datasets with different characteristics. SVM and Roccio classifiers were used for evaluation as the mostly preferred classifiers for evaluating term weighting schemes which are vector-based classifiers such as SVM [13, 28], centroid-based classifiers [19, 42] and *k*NN [30]. It should be noted that Roccio is one of the widely known centroid-based classifiers for text classification. Experimental results indicated that appropriate choice of the term frequency factor can significantly affect the success of all supervised term weighting methods.

The rest of this paper is structured as follows. The main motivation of this paper is explained in the next subsection. The feature selection and term weighting methods utilized in this study are briefly explained in Sects. 2 and 3, respectively. Section 4 describes classifiers employed in the experiments. The dataset characteristics, success measures, accuracy analysis, and statistical analysis are provided in detail in Sect. 5. In Sect. 6, discussions about the results are given. Finally, some concluding remarks and future work are presented in Sect. 7.

1.1 Motivation of the Study

In this study, we aimed to show that the performance of a term weighting scheme is not only based on having an effective collection frequency but also based on selecting suitable term frequency factor.

In the recently proposed term weighting schemes for text classification, a raw term frequency (TF) factor and a newly developed collection frequency factor are generally combined. Using raw TF values in the weighting process complicates the representation of text documents in the vector space, especially when there exist too many higher TF values in the collection. We can explain this problem as follows. Too many higher TF values in the collection can cause that the documents belonging to same class are located away from each other in the vector space, while the documents belonging to different classes are located closer to each other. This situation leads to weaker representation of the documents in the vector space, where the class discrimination potential of the text documents is not fully reflected by the current term weighting methods. Thus, it results in the lower classification performances for the term weighting schemes. To deal with this problem, we employed two modified versions of TF (LOG_TF and SQRT_TF) together with raw TF on seven supervised term weighting schemes and compared their performances.

It should be noted that one of the seven supervised schemes (TF-DFS) used in the experiments is adapted from an existing feature selection method, namely distinguishing feature selector (DFS) [6]. Although this method is known as an effective feature selection method for text classification, it is not adapted as a term weighting method before according to the literature. We employed the weighting scheme, namely TF-DFS, with the above-mentioned three different term frequency factors in the text classification experiments.

We have not only compared the effects of three different term frequency factors on the performances of each term weighting method but also compared the maximum classification performances of seven employed term weighting schemes. Experimental results show that SQRT_TF-DFS scheme generally outperformed other six supervised term weighting schemes having raw TF, LOG_TF, and SQRT_TF.

2 Feature Selection

In this paper, we used a state-of-art filter method, namely distinguishing feature selector (DFS) [6], for feature selection. DFS assigns scores to each term considering their distinguishing power. If DFS assigns a high score for a term, it is stated that related term is a distinctive term. On the contrary, the corresponding term has weak distinguishing power. The formula of DFS is shown in following equation.

$$\mathrm{DFS}(t) = \sum_{i=1}^{M} \frac{P(C_i|t)}{P(\bar{t}|C_i) + P(t|\bar{C}_i) + 1} \tag{1}$$

In this formula, M represents the number of classes in whole collection, $P(C_i|t)$ denotes the conditional probability of class C_i given the presence of term t, $P(\bar{t}|C_i)$ denotes the conditional probability of the term in the absence of term t given class C_i and $P(t|\bar{C}_i)$ denotes the conditional probability of term t given all the classes except C_i, respectively.

3 Term Weighting Schemes

A term weighting scheme generally contains three factors as shown in Eq. 2. These factors are term frequency factor, collection frequency factor, and length normalization factor.

$$\begin{aligned}\text{Weighting scheme } (W) \;=\; & \text{term frequency} \\ & \times \text{ collection frequency} \\ & \times \text{ length normalization}\end{aligned} \tag{2}$$

Most of the newly proposed schemes are focused on developing new collection frequency factors without investigating other parameters very much. In this study, we used three different term frequency factors. Table 1 shows these term frequency factors with their explanations.

We also employed seven supervised weighting schemes to evaluate the performance of various term frequency factors on these schemes. In the supervised term weighting schemes, the class information of a term in different classes of text is used with following expressions in Table 2. This table shows the probable distribution of term t_i related to C_j in the training collection.

Table 1 List of term frequency factors

TF factor	Expression	Explanation
tf	TF	Raw term frequency (number of times a term occurs in a document)
$\log_2(\mathrm{tf}+1)$	LOG_TF	Logarithm of the term frequency to scale the impact of inconveniently high term frequency
sqrt(tf)	SQRT_TF	Square root of the term frequency to scale the impact of inconveniently high term frequency

Table 2 Two-way contingency table of a term t_i and a category C_j for binary classification

	Containing term t_i	Not containing term t_i
Belonging to category C_j	a_{ij}	b_{ij}
Not belonging to category C_j	c_{ij}	d_{ij}

All of these seven term weighting schemes are listed and explained in the next subsections. While some term weighting schemes are derived from feature selection methods, some of them are the methods proposed especially for term weighting. It should be noted that descriptions of feature weighting methods are given using TF as default term frequency factor in the next subsections.

3.1 Term Weighting Based on DFS (TF-DFS)

DFS is originally a feature selection method which is described in the previous sections. It should be noted that DFS is not used as a term weighting method in the literature before. We adapted it for feature weighting by multiplying term frequency factor with the DFS scores. The calculation of DFS weighting depending on expressions in Table 1 is demonstrated in Eq. 3.

$$W_{\mathrm{TF.DFS}}(t_i) = \mathrm{TF}(t_i, d_k) \times \sum_{j=1}^{M} \left(\frac{\left(\frac{a_{ij}}{a_{ij}+c_{ij}}\right)}{\left(\frac{b_{ij}}{a_{ij}+b_{ij}}\right) + \left(\frac{c_{ij}}{c_{ij}+d_{ij}}\right) + 1} \right) \tag{3}$$

In this equation, M is the number of classes in the collection and $\mathrm{TF}(t_i, d_k)$ is the occurrence frequency of term t_i in the document d_k. These two notations are also valid for the following term weighting schemes in this section.

3.2 Term Weighting Based on Chi-Square Statistic (TF-CHI2)

Chi-square is one of the mainly preferred feature selection methods. This approach is a statistical metric which measures the relation between occurrence of specific term t_i and class C_j [5]. For text classification, weighting score of one term-based Chi-square can be calculated as Eq. 4.

$$W_{\mathrm{TF.CHI2}}(t_i) = \mathrm{TF}(t_i, d_k) \times D \times \max_{j=1}^{M} \left\{ \frac{\left(a_{ij} \times d_{ij} - b_{ij} \times c_{ij}\right)^2}{(a_{ij} + c_{ij})(b_{ij} + d_{ij})(a_{ij} + b_{ij})(c_{ij} + d_{ij})} \right\} \quad (4)$$

In this equation, D represents the number of total documents in entire collection. We used maximum globalization function in order to convert class-specific CHI2 values to unique scores for each term.

3.3 Probability-Based Term Weighting (TF-PB)

Probability-based term weighting takes into account intra-class distribution as well as inter-class distribution for a term. It uses the ratios a_{ij}/b_{ij} and a_{ij}/c_{ij} which directly indicate terms' relevance in regard to a specific category [14]. While the ratio a_{ij}/b_{ij} states intra-class distribution of term t_i in category c_j, the ratio a_{ij}/c_{ij} represents inter-class distribution of t_i in category c_j. In Eq. 5, the calculation of probability-based term weighting is shown.

$$W_{\mathrm{TF.PB}}(t_i) = \mathrm{TF}(t_i, d_k) \times \max_{j=1}^{M} \left\{ \log\left(1 + \frac{a_{ij}}{b_{ij}} \times \frac{a_{ij}}{c_{ij}}\right) \right\} \quad (5)$$

In this equation, we used maximum globalization function in order to convert class-specific PB values to unique scores for each term.

3.4 Term Weighting Based on Relevance Frequency (TF-RF)

RF is one of the recently proposed term weighting methods. The idea of this method can be expressed as follows: To efficiently select positive samples from negative samples, it is required to focus on the term which has high frequency in positive category rather than term which has high frequency in negative category [43]. The formula of this weighting approach is demonstrated in Eq. 6.

$$W_{\mathrm{TF.RF}}(t_i) = \mathrm{TF}(t_i, d_k) \times \max_{j=1}^{M} \left\{ \log\left(2 + \frac{a_{ij}}{c_{ij}}\right) \right\} \quad (6)$$

In this equation, we used maximum globalization function in order to convert class-specific RF values to unique scores for each term.

3.5 Term Weighting Based on Inverse Gravity Moment (TF-IGM)

IGM is proposed as a statistical model which is based on calculation of inverse gravity moment of terms. The aim of this model is to obtain term's class distinguishing ability by measuring the concentration or non-uniformity degree of inter-class distribution of a term [27]. IGM value of a term is calculated as Eq. 7.

$$\mathrm{IGM}(t_i) = \frac{f_{i1}}{\sum_{r=1}^{M} f_{ir} \times r} \quad (7)$$

In this equation, $\mathrm{IGM}(t_i)$ represents the inverse gravity moment of the inter-class distribution of term t_i. The frequency, f_{ir} ($r = 1, 2, \ldots, M$), denotes the class-based document frequency of the term, i.e., the number of text documents including the term t_i in the rth category, which are sorted in descending order with the rank r.

The TF-IGM weight of term t_i in document d_k is calculated by multiplying the term frequency factor and the IGM-based collection frequency factor which is expressed as Eq. 8.

$$W_{\mathrm{TF.IGM}}(t_i) = \mathrm{TF}(t_i, d_k) \times (1 + \lambda \times \mathrm{IGM}(t_i)) \quad (8)$$

In this equation, λ is a constant which can be adjusted. The aim of it is to keep the relative balance between the term frequency factor and IGM-based collection frequency factors in weight of a term. In the referenced paper, the λ coefficient's value range is determined as 5.0–9.0 and the default value of λ is determined as 7.0. In this study, we used the default value for λ coefficient in the experiments.

3.6 Term Weighting Based on Inverse Class Frequency (TF-IDF-ICF)

This supervised weighting scheme is created by adding inverse class frequency information of terms to traditional TF-IDF. The characteristic of TF.IDF.ICF term weighting scheme is that it favors the rare terms and it is biased against frequent terms [19]. The weighting formula of TF-IDF-ICF is shown in Eq. 9.

$$W_{\mathrm{TF.IDF.ICF}}(t_i) = \mathrm{TF}(t_i, d_k) \times \left(1 + \log\left(\frac{D}{d(t_i)}\right)\right) \times \left(1 + \log\left(\frac{C}{c(t_i)}\right)\right) \quad (9)$$

In this equation, C is the total number of predefined classes in collection and $c(t_i)$ is the number of classes in the collection in which term t_i occurs at least once. However, $D/d(t_i)$ and $C/c(t_i)$ represent class frequency (CF) and ICF of term t_i, respectively.

3.7 Term Weighting Based on Inverse Class Density Frequency (TF-IDF-ICSDF)

This term weighting scheme weights terms by calculating their inverse class space density frequency (ICSDF) value. The idea of TF.IDF.ICSDF term weighting is that it provides a positive discrimination for both frequent and rare terms, and it is effective in both low-dimensional and high-dimensional vector spaces [19]. The weighting formula of TF.IDF.ICSDF is shown in the equation below.

$$W_{\text{TF.IDF.ICSDF}}(t_i) = \text{TF}(t_i, d_k) \times \left(1 + \log\left(\frac{D}{d(t_i)}\right)\right) \times \left(1 + \log\left(\frac{C}{\sum_{j=1}^{M} \frac{df_{t_{ij}}}{D_j}}\right)\right) \tag{10}$$

In this equation, $df_{t_{ij}}$ and D_j are the document frequencies of term t_i for the jth class and the total number of text documents occurring in class C_j (j=1, 2, …, M), respectively.

4 Classifiers

In text classification, researchers have benefited from lots of classifiers up to now on the learning process of documents. The mostly preferred classifiers for evaluating term weighting schemes are vector-based classifiers such as SVM [13, 28], centroid-based classifiers [19, 42], and kNN [30]. We used SVM and Roccio classifier [44] which is derived from centroid-based classifiers in our experiments to evaluate the performance of term weighting schemes.

4.1 Support Vector Machines (SVM)

SVM is a learning algorithm separating positive samples from negative samples by creating a linear or nonlinear hyperplane. The location of the hyperplane in hyperspace is the maximizing distance between nearest positive and negative samples named as support vectors. The distance from the hyperplane to the closest data point determines the margin of the classifier. Thus, the essential point of SVM classifier is the maximization of the margin [7]. SVM can be a linear or nonlinear classifier depending on the selection of the kernel parameter.

In this study, we preferred linear SVM classifier because it generally shows good performance for text classification. We used LibSVM [45] with default parameter settings in the experiments. It is a popular open-source software package that supports multiclass classification tasks.

4.2 Roccio

Roccio is a classification algorithm derived from centroid classification which is commonly used for text classification. In vector space model (VSM), each document vector d_j (j=1, 2, …, n) belonging to class C_k (k=1, 2, …, m) is represented by a multi-dimensional feature vector. The calculation of centroid of a certain class C_k is done according to the following equations:

Firstly, the sum of the training document vectors in the same class is calculated as in Eq. 11.

$$C_k^{\text{sum}} = \sum_{d \in c_k} d_j \tag{11}$$

Then, centroid vectors for each class are calculated via dividing these sums by the number of training documents in the corresponding class as in Eq. 12.

$$\text{Centroid}_k = C_k^{\text{sum}} / n_{C_k} \tag{12}$$

The distance between each test document t_s (s=1, 2, …, r) and all centroids Centroid_k is calculated by the following formula.

$$D_{sk} = \sqrt{(t_s - \text{Centroid}_k)^2} \tag{13}$$

Finally, the test document t_s is assigned to the class of the corresponding centroid having minimum D_{sk} value. This operation is repeated for all test documents.

5 Experimental Work

An in-depth investigation was realized to measure the impacts of various term frequency factors on supervised term weighting schemes using different classifiers. The general classification procedure explaining the experimental study is summarized in Fig. 1.

In order to validate the performance of term frequency factors, two different datasets with varying characteristics were utilized. We performed lowercase conversion, removed stop-words, and carried out stemming [46] on the three-stage preprocessing process. Porter stemming algorithm was utilized for obtaining stemmed forms of the words. DFS is used as the feature selection method, and experiments were realized with different feature sizes ranging from 300 to 4000.

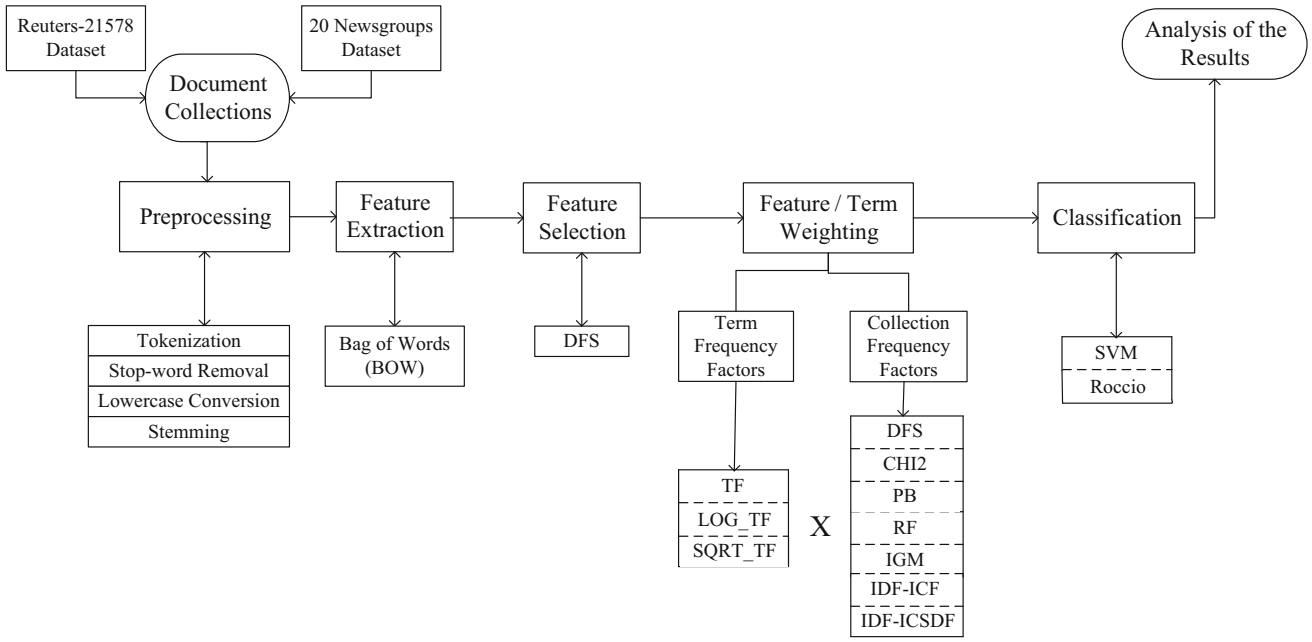

Fig. 1 General classification procedure in the experimental study

The used datasets and success measure are shortly explained in the next subsections. Then, the experimental results and statistical analyses are presented.

5.1 Datasets

In the experiments, we used two benchmark datasets, namely Reuters-21578 and 20 Newsgroups, to analyze the effects of term frequency factor on the term weighting schemes. The former dataset contains top-10 classes of well-recognized Reuters-21578 ModApte split [47] which is employed in many studies in the literature [7]. The latter dataset consists of ten classes of another well-known text collection, namely 20 Newsgroups [47]. Reuters-21578 is an imbalanced dataset where there exists different number of documents for each class. 20 Newsgroups is a widely used balanced benchmark dataset where the number of text documents in individual classes is equal. While Reuters-21578 dataset has its own training and testing split, 20 Newsgroups dataset was manually divided into training and testing splits. For this purpose, 20 Newsgroups dataset was divided into two equal parts (50% and 50%) for training and testing. Tables 3 and 4 show the extensive information related to utilized datasets.

5.2 Success Measure

In text classification, metrics based on precision (P) and recall (R) have been widely preferred for measuring performance of classification. Assume that FP, TP, and FN denote the false positives, true positives, and false negatives, respec-

Table 3 Reuters dataset

No.	Class label	Training documents	Testing documents
1	earn	2877	1087
2	acq	1650	719
3	money-fx	538	179
4	grain	433	149
5	crude	389	189
6	trade	369	117
7	interest	347	131
8	ship	197	89
9	wheat	212	71
10	corn	181	56

Table 4 Newsgroups dataset

No.	Class label	Training documents	Testing documents
1	alt.atheism	500	500
2	comp.graphics	500	500
3	comp.os.ms-windows.misc	500	500
4	comp.sys.ibm.pc.hardware	500	500
5	comp.sys.mac.hardware	500	500
6	comp.windows.x	500	500
7	misc.forsale	500	500
8	rec.autos	500	500
9	rec.motorcycles	500	500
10	rec.sport.baseball	500	500

Table 5 Micro-F_1 scores (%) obtained on Reuters-21578 dataset using TF-DFS term weighting scheme with three different term frequency factors

	SVM			Roccio		
	TF	LOG_TF	SQRT_TF	TF	LOG_TF	SQRT_TF
300	86.69	87.12	**87.44**	74.92	81.45	**81.74**
500	86.62	87.30	**87.33**	75.46	81.52	**81.92**
1000	86.33	87.51	**87.62**	75.67	81.63	**82.02**
1500	87.12	**87.80**	87.69	75.75	81.63	**82.17**
2000	86.98	87.77	***87.84***	75.85	81.81	***82.20***
3000	87.23	**87.77**	87.73	75.89	81.63	**82.13**
4000	87.30	***87.84***	87.69	75.92	81.67	**82.17**

Table 6 Micro-F_1 scores (%) obtained on Reuters-21578 dataset using TF-CHI2 term weighting scheme with three different term frequency factors

	SVM			Roccio		
	TF	LOG_TF	SQRT_TF	TF	LOG_TF	SQRT_TF
300	76.89	84.64	**84.93**	72.87	76.89	**77.93**
500	84.82	85.22	**85.33**	72.87	76.89	**77.97**
1000	85.47	86.55	***86.87***	72.87	76.89	**78.01**
1500	85.25	86.11	**86.19**	72.87	76.89	**78.01**
2000	85.22	86.19	**86.40**	72.87	76.89	**78.01**
3000	85.72	**86.47**	86.40	72.87	76.89	**78.01**
4000	85.58	86.47	**86.47**	72.87	76.89	***78.01***

Table 7 Micro-F_1 scores (%) obtained on Reuters-21578 dataset using TF-PB term weighting scheme with three different term frequency factors

	SVM			Roccio		
	TF	LOG_TF	SQRT_TF	TF	LOG_TF	SQRT_TF
300	85.94	**86.11**	85.79	70.97	74.63	**75.42**
500	86.04	**86.26**	86.01	70.97	74.67	**75.49**
1000	86.04	***86.33***	86.01	70.72	75.03	**75.49**
1500	86.08	**86.29**	85.97	70.72	75.03	**75.49**
2000	86.08	**86.29**	85.97	70.72	75.03	**75.49**
3000	86.08	**86.29**	85.97	70.72	75.03	**75.49**
4000	86.08	**86.29**	85.97	70.72	75.03	***75.49***

Table 8 Micro-F_1 scores (%) obtained on Reuters-21578 dataset using TF-RF term weighting scheme with three different term frequency factors

	SVM			Roccio		
	TF	LOG_TF	SQRT_TF	TF	LOG_TF	SQRT_TF
300	86.08	**86.87**	86.80	71.26	77.43	**78.08**
500	86.40	86.51	**86.62**	71.33	77.79	**78.22**
1000	86.44	86.76	**86.87**	71.44	77.83	**78.40**
1500	86.55	***87.23***	***87.23***	71.44	77.93	***78.58***
2000	86.44	86.98	**87.01**	71.44	78.08	**78.54**
3000	86.62	87.01	**87.08**	71.44	78.08	**78.47**
4000	86.87	**87.05**	86.98	71.48	78.08	**78.47**

tively. Then, precision and recall can be expressed as in Eq. 14.

$$P = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})\, R = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) \tag{14}$$

F_1 measure incorporates both precision and recall scores. Since these metrics (P, R, and F_1) are special to binary classification, micro-averaged F_1 (micro-F_1) and macro-averaged F_1 (macro-F_1) metrics are used for multiclass classification problems. We used micro-F_1 metric which is shown with the following definition in our experiments.

$$\text{Micro}-F_1 = 2 \times (P \times R)/(P+R) \tag{15}$$

5.3 Accuracy Analysis

In this section, we analyzed the performance of classification with SVM and Roccio classifiers on the Reuters-21578 and 20 Newsgroups datasets using various feature sizes. All of the experiments were realized using three different term frequency factors in seven supervised term weighting schemes. These term frequency factors are TF, SQRT_TF, and LOG_TF. TF is the default term frequency factor, and the others are modified versions of TF term frequency factor.

Tables 5, 6, 7, 8, 9, 10 and 11 show the micro-F_1 scores obtained on Reuters-21578 dataset with seven different term weighting schemes using SVM and Roccio classifiers. We showed maximum scores of term weighting methods in the tables as bolded. Also, maximum score of each term weighting scheme based on classifier is shown as bold italic.

One can note that almost all modified schemes (SQRT_TF and LOG_TF) outperform the TF term frequency factor for all feature dimensions. Although LOG_TF term frequency factor is more popular than SQRT_TF term frequency factor for term weighting studies, the performance of SQRT_TF term frequency factor is superior to the performance of LOG_TF term frequency factor in general. Also, TF-DFS generally obtains better results than the other schemes on two classifiers.

For SVM classifier, the micro-F_1 scores are generally high (mostly bigger than 85%) for all feature sizes and the highest micro-F_1 score (87.84%) is attained by the LOG_TF-DFS and SQRT_TF-DFS term weighting schemes. While the lowest micro-F_1 score (76.9%) is obtained with TF-CHI2 weighting scheme using 300 features for SVM classifier, micro-F_1 scores obtained with SQRT_TF-CHI2 and

Table 9 Micro-F_1 scores (%) obtained on Reuters-21578 dataset using TF-IGM term weighting scheme with three different term frequency factors

	SVM			Roccio		
	TF	LOG_TF	SQRT_TF	TF	LOG_TF	SQRT_TF
300	85.54	**86.76**	86.58	75.64	79.98	**80.55**
500	85.76	**86.44**	86.37	76.39	80.91	**81.27**
1000	86.04	86.87	**87.05**	76.86	81.45	**81.77**
1500	86.01	87.12	**87.23**	77.40	81.63	**81.81**
2000	85.83	87.08	***87.30***	77.47	81.74	**81.84**
3000	85.94	87.08	**87.23**	77.47	81.77	**81.88**
4000	86.08	86.94	**87.23**	77.68	81.88	***82.13***

Table 10 Micro-F_1 scores (%) obtained on Reuters-21578 dataset using TF-IDF-ICF term weighting scheme with three different term frequency factors

	SVM			Roccio		
	TF	LOG_TF	SQRT_TF	TF	LOG_TF	SQRT_TF
300	85.76	86.62	**86.87**	78.90	81.59	***82.17***
500	85.72	86.22	**86.40**	79.37	81.38	**81.95**
1000	85.76	86.26	**86.58**	80.09	81.74	**82.17**
1500	85.94	86.80	***86.98***	80.66	81.92	**82.13**
2000	85.61	**86.69**	86.58	80.70	**81.92**	**81.92**
3000	85.61	86.44	**86.76**	80.88	**81.95**	81.77
4000	85.54	**86.69**	**86.69**	80.95	81.74	**81.74**

Table 11 Micro-F_1 scores (%) obtained on Reuters-21578 dataset using TF-IDF-ICSDF term weighting scheme with three different term frequency factors

	SVM			Roccio		
	TF	LOG_TF	SQRT_TF	TF	LOG_TF	SQRT_TF
300	85.25	86.40	***86.47***	80.19	82.78	***83.14***
500	84.97	85.15	**85.68**	80.34	82.24	**82.53**
1000	85.47	85.83	**86.01**	80.80	**81.63**	81.06
1500	85.33	85.94	**86.01**	81.02	**81.31**	81.09
2000	85.00	85.47	**85.65**	80.41	**80.91**	80.41
3000	85.25	85.68	**85.76**	79.55	**80.45**	80.30
4000	85.25	85.94	**85.97**	79.55	**80.52**	80.05

LOG_TF-CHI2 weighting schemes using same amount of features are higher than 84.6%. So, it can be said that TF term frequency factor is the reason behind this unsuccessful result. In other words, modification on TF term frequency factor affected the performance of classification positively even using less number of features for SVM classifier.

Higher performance difference (almost 7%) between TF and modified versions of TF term frequency factor is achieved with Roccio classifier. It should be noted that micro-F_1 scores

Table 12 Micro-F_1 scores (%) obtained on 20 Newsgroups dataset using TF-DFS term weighting scheme with three different term frequency factors

	SVM			Roccio		
	TF	LOG_TF	SQRT_TF	TF	LOG_TF	SQRT_TF
300	97.36	98.30	**98.42**	88.94	97.88	**98.24**
500	97.34	98.04	**98.22**	88.70	97.90	***98.26***
1000	97.36	98.26	***98.54***	89.20	97.76	**98.18**
1500	97.26	98.32	**98.50**	89.20	97.78	**98.18**
2000	97.44	98.30	**98.44**	89.34	97.80	**98.16**
3000	97.48	98.30	**98.44**	89.46	97.90	**98.20**
4000	97.40	98.26	**98.38**	89.50	97.86	**98.18**

Table 13 Micro-F_1 scores (%) obtained on 20 Newsgroups dataset using TF-CHI2 term weighting scheme with three different term frequency factors

	SVM			Roccio		
	TF	LOG_TF	SQRT_TF	TF	LOG_TF	SQRT_TF
300	96.34	97.74	**97.90**	84.00	96.50	**96.94**
500	96.82	**97.98**	97.94	84.00	96.52	**96.94**
1000	97.08	97.84	***98.10***	83.72	96.48	**96.96**
1500	97.30	97.88	**98.00**	83.72	96.48	**96.96**
2000	97.38	97.84	**98.04**	83.72	96.48	**96.96**
3000	97.32	97.82	**98.06**	83.72	96.48	**96.96**
4000	97.38	97.82	**98.06**	83.72	96.48	***96.96***

of Roccio classifier do not change very much depending on the feature size such as SVM classifier. While the lowest micro-F_1 score for Roccio classifier is 70.72% with TF-PB term weighting scheme using 1000 features, the highest micro-F_1 score is 83.14% with SQRT_TF-IDF-ICSDF term weighting scheme using 300 features.

Tables 12, 13, 14, 15, 16, 17 and 18 show the micro-F_1 scores obtained on 20 Newsgroups dataset with seven different term weighting schemes using SVM and Roccio classifiers. As in Reuters-21578 dataset, term weighting schemes with modified term frequency factors (SQRT_TF and LOG_TF) generally outperform the ones with TF term frequency factor. The performance of TF-PB weighting scheme with either TF term frequency factor or its modified versions is worse than the other term weighting schemes on 20 Newsgroups dataset. It is necessary to note that TF-PB weighting scheme is known as proper for the classification of imbalanced datasets rather than balanced ones. This reason may cause a decrease in the performance of classification on 20 Newsgroups dataset. It should be noted that the performance of TF-IDF-ICF and TF-IDF-ICSDF with TF term frequency factor decreased dramatically as feature size increases. However, the performances of the same

Table 14 Micro-F_1 scores (%) obtained on 20 Newsgroups dataset using TF-PB term weighting scheme with three different term frequency factors

	SVM			Roccio		
	TF	LOG_TF	SQRT_TF	TF	LOG_TF	SQRT_TF
300	**77.34**	77.20	77.12	49.02	55.44	**58.54**
500	***77.94***	77.90	77.28	49.04	55.48	***58.56***
1000	**77.64**	76.84	76.16	41.00	50.78	**53.64**
1500	**77.76**	77.18	76.30	41.00	50.78	**53.70**
2000	**77.76**	77.32	76.32	41.00	50.78	**53.68**
3000	**77.68**	77.30	76.36	41.00	50.78	**53.70**
4000	**77.84**	77.32	76.38	41.00	50.78	**53.68**

Table 15 Micro-F_1 scores (%) obtained on 20 Newsgroups dataset using TF-RF term weighting scheme with three different term frequency factors

	SVM			Roccio		
	TF	LOG_TF	SQRT_TF	TF	LOG_TF	SQRT_TF
300	96.82	98.02	**98.16**	61.82	69.34	**73.90**
500	96.58	97.76	**98.02**	62.10	69.84	***74.48***
1000	96.80	**98.18**	**98.18**	46.04	61.30	**66.70**
1500	97.04	98.32	**98.38**	46.38	61.60	**66.86**
2000	97.22	98.28	***98.40***	46.60	62.02	**67.08**
3000	97.12	**98.30**	**98.30**	46.84	62.24	**67.28**
4000	97.28	98.22	**98.30**	46.86	62.42	**67.42**

Table 16 Micro-F_1 scores (%) obtained on 20 Newsgroups dataset using TF-IGM term weighting scheme with three different term frequency factors

	SVM			Roccio		
	TF	LOG_TF	SQRT_TF	TF	LOG_TF	SQRT_TF
300	96.68	97.74	**97.90**	87.94	96.78	***97.46***
500	97.14	**97.90**	**97.90**	88.10	96.82	**97.30**
1000	97.32	98.02	***98.18***	88.38	96.72	**97.24**
1500	97.24	97.88	**98.08**	88.02	96.70	**97.12**
2000	96.98	98.00	**98.08**	88.20	96.62	**97.02**
3000	97.10	98.00	**98.12**	88.22	96.54	**96.98**
4000	97.10	97.98	**98.14**	88.16	96.56	**96.94**

schemes with modified term frequency factors (SQRT_TF and LOG_TF) do not have such a decrease depending on feature size. So, it can be said that modified term frequency factors make these schemes more stable.

Micro-F_1 scores show that all weighting schemes except TF-PB obtained higher performances with modified term frequency factors for all feature sizes using SVM classifier. The maximum rise in the performance of classification is achieved as 3.7% on TF-IDF-ICSDF term weighting scheme

Table 17 Micro-F_1 scores (%) obtained on 20 Newsgroups dataset using TF-IDF-ICF term weighting scheme with three different term frequency factors

	SVM			Roccio		
	TF	LOG_TF	SQRT_TF	TF	LOG_TF	SQRT_TF
300	96.78	97.86	**98.12**	81.92	92.66	***93.02***
500	96.46	97.82	**97.92**	80.34	92.36	**92.90**
1000	96.64	98.04	**98.16**	78.76	92.04	**92.52**
1500	96.30	98.02	***98.18***	78.50	91.78	**92.16**
2000	96.40	97.88	**98.02**	78.48	91.74	**91.90**
3000	96.10	**98.06**	**98.06**	78.44	91.40	**91.76**
4000	96.06	**97.98**	**97.98**	78.36	91.30	**91.64**

Table 18 Micro-F_1 scores (%) obtained on 20 Newsgroups dataset using TF-IDF-ICSDF term weighting scheme with three different term frequency factors

	SVM			Roccio		
	TF	LOG_TF	SQRT_TF	TF	LOG_TF	SQRT_TF
300	96.04	97.50	**97.76**	78.02	89.62	***90.28***
500	95.94	97.40	**97.60**	76.14	89.66	**90.00**
1000	95.48	97.72	***97.92***	74.28	89.14	**89.50**
1500	95.12	97.66	**97.90**	74.02	**88.70**	**88.70**
2000	94.72	97.48	**97.74**	74.12	88.14	**88.50**
3000	94.00	96.94	**97.26**	74.30	87.46	**87.82**
4000	93.42	96.82	**97.12**	73.76	87.14	**87.38**

with 4000 features. For SVM classifier, the performance of SQRT_TF term frequency factor is generally better than the others apart from TF-PB term weighting scheme.

Higher performance difference (20.67%) between SQRT_TF-RF (66.7%) and TF-RF (46.04%) is achieved with Roccio classifier on 20 Newsgroups dataset. Increasing feature size does not affect micro-F_1 scores of Roccio classifier very much. While the lowest micro-F_1 score on 20 Newsgroups dataset with Roccio classifier is obtained as (41%) with TF-PB term weighting scheme using 1000 or more features, the highest micro-F_1 score is obtained as (98.26%) with SQRT_TF-DFS term weighting scheme using 500 features.

According to the results obtained on both datasets, it can be said that the range of the performance change in Reuters-21578 is lower than the ones obtained in 20 Newsgroups dataset. It may be caused from the class distributions of documents in each dataset. While 20 Newsgroups dataset is an example for balanced datasets, Reuters-21578 dataset is highly skewed. Also, term weighting schemes using SQRT_TF term frequency factor are more successful than the ones using TF and LOG_TF term frequency factors in general. TF term frequency factor seems as the least effective

Table 19 Maximum micro-F_1 scores (%) obtained on seven term weighting schemes with three different term frequency factors

Datasets	Classifiers	TF-DFS	TF-CHI2	TF-PB	TF-RF	TF-IGM	TF-IDF-ICF	TF-IDF-ICSDF
Reuters-21578	SVM	**87.84** (SQRT_TF)	86.87 (SQRT_TF)	86.33 (LOG_TF)	87.23 (SQRT_TF)	87.30 (SQRT_TF)	86.98 (SQRT_TF)	86.47 (SQRT_TF)
	Roccio	82.20 (SQRT_TF)	78.01 (SQRT_TF)	75.49 (SQRT_TF)	78.58 (SQRT_TF)	82.13 (SQRT_TF)	82.17 (SQRT_TF)	**83.14** (SQRT_TF)
20 Newsgroups	SVM	**98.54** (SQRT_TF)	98.10 (SQRT_TF)	77.94 (TF)	98.40 (SQRT_TF)	98.18 (SQRT_TF)	98.18 (SQRT_TF)	97.92 (SQRT_TF)
	Roccio	**98.26** (SQRT_TF)	96.96 (SQRT_TF)	58.56 (SQRT_TF)	74.48 (SQRT_TF)	97.46 (SQRT_TF)	93.02 (SQRT_TF)	90.28 (SQRT_TF)

Table 20 Statistical significance of performance improvements in modified term frequency factors in Reuters-21578 dataset

Weighting scheme	SVM		Roccio	
	TF versus LOG_TF	TF versus SQRT_TF	TF versus LOG_TF	TF versus SQRT_TF
TF-DFS	0.0002*	0.0003*	0.0000*	0.0000*
TFCHI2	0.0587***	0.0525***	0.0000*	0.0000*
TF-PB	0.0000*	0.0008*	0.0000*	0.0000*
TF-RF	0.0021*	0.0009*	0.0000*	0.0000*
TF-IGM	0.0000*	0.0000*	0.0000*	0.0000*
TF-IDF-ICF	0.0000*	0.0000*	0.0004*	0.0012*
TF-IDF-ICSDF	0.0016*	0.0000*	0.0051*	0.0346**

*Significance at 99%
**Significance at 95%
***Significance at 90%

Table 21 Statistical significance of performance improvements in modified term frequency factors in 20 Newsgroups dataset

Weighting scheme	SVM		Roccio	
	TF versus LOG_TF	TF versus SQRT_TF	TF versus LOG_TF	TF versus SQRT_TF
TF-DFS	0.0000*	0.0000*	0.0000*	0.0000*
TFCHI2	0.0009*	0.0002*	0.0000*	0.0000*
TF-PB	0.0028*	0.0005*	0.0000*	0.0000*
TF-RF	0.0000*	0.0000*	0.0000*	0.0000*
TF-IGM	0.0000*	0.0000*	0.0000*	0.0000*
TF-IDF-ICF	0.0000*	0.0000*	0.0000*	0.0000*
TF-IDF-ICSDF	0.0000*	0.0000*	0.0000*	0.0000*

*Significance at 99%

one among three different term frequency factors according to the experimental results.

Table 19 shows the maximum classification performances obtained on seven term weighting schemes with three different term frequency factors. The expressions in the brackets under the micro-F_1 values in the table demonstrate the term frequency factor used in the corresponding term weighting scheme. For example, the maximum performance score of TF-DFS is acquired as 87.84 with SQRT_TF term frequency factor for SVM classifier on the Reuters-21578 dataset.

The results show that TF-DFS which is adapted from a feature selection method for the first time generally outperformed other six supervised term weighting schemes. Moreover, SQRT_TF-DFS has promising results over recent term weighting methods for text classification in the literature such as TF-IGM, TF-IDF-ICF, and TF-IDF-ICSDF.

5.4 Statistical Analysis

To demonstrate the validity of each modified term frequency factor, we also used the t test. Tables 20 and 21 show the results of P values from one-tailed paired t-test. The results showed that the performance gains obtained by modified term frequency factors (LOG_TF and SQRT_TF) compared to raw

Table 22 Sample collection including features and their frequencies

Class 1	Cat	Dog	Mouse	Class 2	Cat	Dog	Mouse
Doc1	43	20	1	Doc9	20	3	4
Doc2	30	10	2	Doc10	30	2	8
Doc3	51	14	3	Doc11	36	6	5
Doc4	49	15	5	Doc12	29	8	5
Doc5	47	13	4	Doc13	27	6	9
Doc6	40	13	2	Doc14	32	5	4
Doc7	41	14	7	Doc15	25	5	10
Doc8	42	17	6	Doc16	30	5	7

term frequency factor are statistically significant with a rather high confidence with 90% and 99% levels for Reuters-21578 and 20 Newsgroups datasets, respectively. As seen from the tables, almost all *P-values* except a few cases are at 99% confidence level. These results verify the superiority of modified term frequency factors against raw term frequency factor in term weighting schemes.

6 Discussions

In this section, in order to discuss the results of the experimental analysis, a sample collection is provided in Table 22. In text classification, text documents are represented with multi-dimensional feature vectors. However, visualization of a multi-dimensional feature vector whose number of dimensions is more than 3 is very difficult. Therefore, the sample collection provided in the table consists of three features. There exist 16 documents in the collection belonging to two classes. The term frequencies of three features are shown in Table 22.

In the three-dimensional space, the document vectors based only on term frequency factors of these three features are shown in Fig. 2.

It should be noted that it seems very difficult to classify all document vectors located in Fig. 2 correctly because the document vectors of different classes are nested. A good term weighting scheme must adjust the positions of text documents in the vector space in order to improve the classifier's performance. Let us assume that the scores, term weighting scheme named X, produced for 'Cat,' 'Dog,' and 'Mouse' features are 1.76, 3.91, and 4.1, respectively. The positions of document vectors in the vector space with these feature weights are demonstrated in Fig. 3.

According to Fig. 3, a classifier may better distinguish document vectors of two classes after weighting by the X method. As seen in this example, only using term frequency factor may not be enough to increase the performance of the classifier in most cases because of high frequency factors

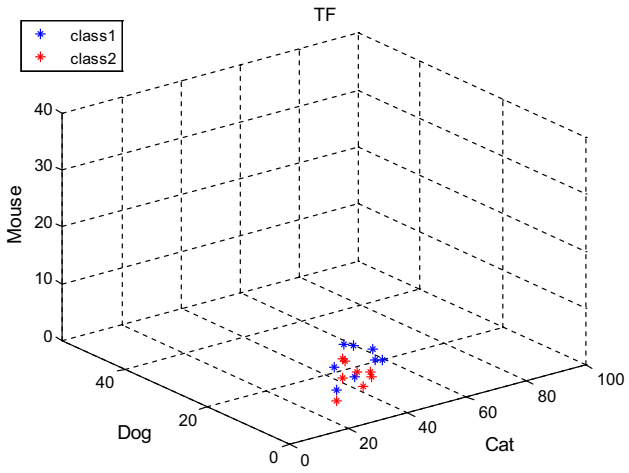

Fig. 2 Document vectors in three-dimensional space

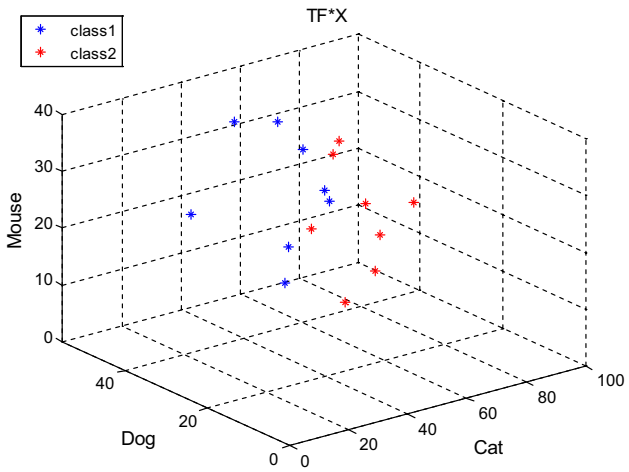

Fig. 3 Weighted document vectors in three-dimensional space

of features. In order to show the effect of using modified term frequency factors, documents vectors weighted with LOG_TF-X and SQRT_TF-X weighting schemes are also presented in Fig. 4 using three-dimensional space.

As shown in Fig. 4a, b, the reduction in the effect of the high term frequencies by applying modified term frequency factors may cause a better discrimination among classes. In this example, the document vectors of the same class seem located closer to each other and this situation may facilitate classification process.

In conclusion, we tried to illustrate how the document vectors' points position in vector space model when a term weighting scheme using raw TF, LOG_TF, and SQRT_TF is employed in the weighting process of terms. For this aim, the distributions of raw, logarithmic, and squared root forms of term frequency factors in vector space model are shown, respectively. As shown in Fig. 3, when the term frequencies are high, it may not be an effective selection using raw term frequencies since it is not fully reflecting the actual ability of

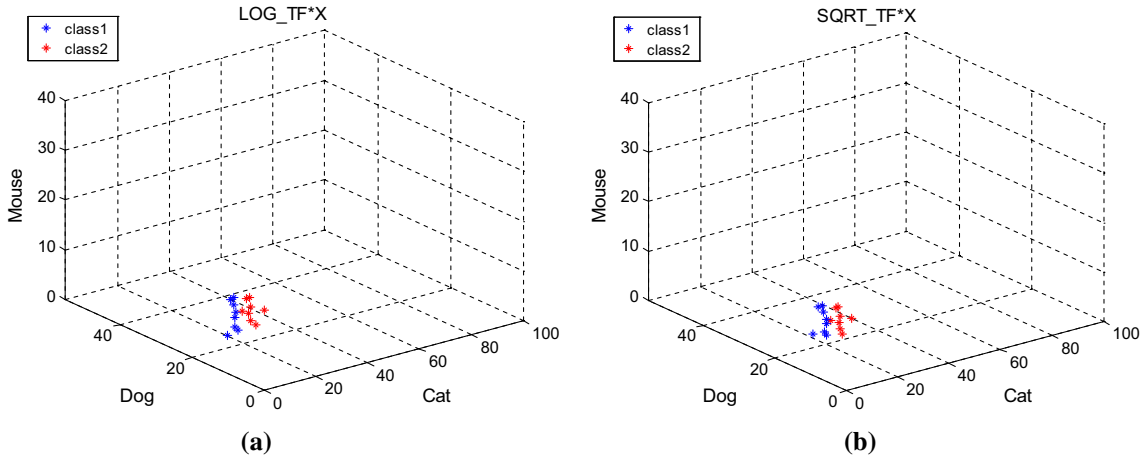

Fig. 4 Representation of these document vectors using LOG_TF (**a**) and SQRT_TF (**b**) weighting schemes in three-dimensional space

existing term weighting schemes. Figure 4 shows that reducing the effects of high TF values with modified LOG_TF and SQRT_TF term frequency factors is an effective solution to provide better representations of document vectors in vector space model. As it is supported with the results presented in Sect. 5.3, these representations caused more successful classification results for term weighting schemes.

7 Conclusions and Future Work

In this study, we extensively analyzed the effects of modifications on term frequency factor of supervised term weighting schemes. We performed this analysis using three different term frequency factors, seven supervised term weighting schemes, two widely known classifiers, and two benchmark datasets with different characteristics. SVM and Roccio were utilized as classification algorithms as they are widely known vector-based classifiers. Appropriate choice of term frequency factor in supervised term weighting schemes may cause that term weighting schemes and classifiers accurately reflect their performance for text classification. In particular, high term frequency values need to be adjusted in order to represent document vectors in vector space model better. To deal with this challenge, the effect of high term frequency factor in supervised term weighting schemes must be reduced appropriately. Experimental results showed that modification of term frequency factor used in supervised term weighting schemes increased the performance of almost all weighting schemes on both balanced and imbalanced text collections. If term frequency factor is modified properly, the success of weighting scheme can be influenced positively because of adjusting the location of intra-class document vectors of a specific class. So, this may facilitate classification process and enhance the performance of classification. According to the experimental findings, although LOG_TF term frequency factor is more popular than SQRT_TF term frequency factor for term weighting studies, the performance of SQRT_TF term frequency factor is superior to the performance of LOG_TF term frequency factor in general. Shortly, we tried to show that the general classification performance of existing weighting scheme may be influenced significantly with not only using a newly proposed collection frequency factor but also selection of suitable term frequency factor according to dataset under concern. Experimental results supported the idea that transformation using logarithmic and square root function of term frequency values in term weighting schemes may lead to better performance results compared to the raw form of them.

It should be also noted that TF-DFS term weighting scheme used in the experiments is adapted from an existing feature selection method, namely distinguishing feature selector. Although this method is known as an effective feature selection method for text classification, it is not adapted as a term weighting method before according to the literature. Experimental results show that SQRT_TF-DFS scheme generally outperformed other six supervised term weighting schemes having raw TF, LOG_TF, and SQRT_TF. In the future, we will investigate for finding a new term frequency factor which can at least obtain a competitive performance with the existing ones. Besides, we aim to develop a novel term weighting method comparable with existing term weighting methods using the more appropriate term frequency factor.

Compliance with Ethical Standards

Conflict of interest This manuscript is the original work of the author and has not been published nor has it been submitted simultaneously elsewhere. It is to specifically state that no competing interests are at stake, and there is no conflict of interest with other people or organi-

zations that could inappropriately influence or bias the content of the paper.

References

1. Uysal, A.K.; Gunal, S.: The impact of preprocessing on text classification. Inf. Process. Manag. **50**(1), 104–112 (2014)
2. Schneider, K.-M.: Weighted average pointwise mutual information for feature selection in text categorization. In: European Conference on Principles of Data Mining and Knowledge Discovery, pp. 252–263. Springer (2005)
3. Lee, C.; Lee, G.G.: Information gain and divergence-based feature selection for machine learning-based text categorization. Inf. Process. Manag. **42**(1), 155–165 (2006). https://doi.org/10.1016/j.ipm.2004.08.006
4. Ogura, H.; Amano, H.; Kondo, M.: Feature selection with a measure of deviations from Poisson in text categorization. Expert Syst. Appl. **36**(3), 6826–6832 (2009). https://doi.org/10.1016/j.eswa.2008.08.006
5. Chen, Y.-T.; Chen, M.C.: Using Chi square statistics to measure similarities for text categorization. Expert Syst. Appl. **38**(4), 3085–3090 (2011). https://doi.org/10.1016/j.eswa.2010.08.100
6. Uysal, A.K.; Gunal, S.: A novel probabilistic feature selection method for text classification. Knowl. Based Syst. **36**, 226–235 (2012). https://doi.org/10.1016/j.knosys.2012.06.005
7. Uysal, A.K.: An improved global feature selection scheme for text classification. Expert Syst. Appl. **43**, 82–92 (2016). https://doi.org/10.1016/j.eswa.2015.08.050
8. Deng, Z.-H.; Tang, S.-W.; Yang, D.-Q.; Zhang, M.; Li, L.-Y.; Xie, K.Q.: A comparative study on feature weight in text categorization. In: APWeb, pp. 588–597. Springer (2004)
9. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. (CSUR) **34**(1), 1–47 (2002)
10. Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. J. Doc. **28**(1), 11–21 (2004). https://doi.org/10.1108/eb026526
11. Debole, F; Sebastiani, F.: Supervised term weighting for automated text categorization. In: Text Mining and its Applications, pp. 81–97. Springer (2004)
12. Lertnattee, V.; Theeramunkong, T.: Analysis of inverse class frequency in centroid-based text classification. In: IEEE International Symposium on Communications and Information Technology, 2004. ISCIT 2004, pp. 1171–1176. IEEE (2004)
13. Lan, M.; Tan, C.L.; Su, J.; Lu, Y.: Supervised and traditional term weighting methods for automatic text categorization. IEEE Trans. Pattern Anal. Mach. Intell. **31**(4), 721–735 (2009)
14. Liu, Y.; Loh, H.T.; Sun, A.: Imbalanced text classification: a term weighting approach. Expert Syst. Appl. **36**(1), 690–701 (2009). https://doi.org/10.1016/j.eswa.2007.10.042
15. Altınçay, H.; Erenel, Z.: Analytical evaluation of term weighting schemes for text categorization. Pattern Recognit. Lett. **31**(11), 1310–1323 (2010). https://doi.org/10.1016/j.patrec.2010.03.012
16. Deisy, C.; Gowri, M.; Baskar, S.; Kalaiarasi, S.; Ramraj, N.: A novel term weighting scheme MIDF for text categorization. J. Eng. Sci. Technol. **5**(1), 94–107 (2010)
17. Wei, B.; Feng, B.; He, F.; Fu, X.: An extended supervised term weighting method for text categorization. In: Proceedings of the International Conference on Human-centric Computing 2011 and Embedded and Multimedia Computing 2011. Lecture Notes in Electrical Engineering, pp. 87–99. (2011). https://doi.org/10.1007/978-94-007-2105-0_11
18. Luo, Q.; Chen, E.; Xiong, H.: A semantic term weighting scheme for text categorization. Expert Syst. Appl. **38**(10), 12708–12716 (2011). https://doi.org/10.1016/j.eswa.2011.04.058
19. Ren, F.; Sohrab, M.G.: Class-indexing-based term weighting for automatic text classification. Inf. Sci. **236**, 109–125 (2013). https://doi.org/10.1016/j.ins.2013.02.029
20. Emmanuel, M.; Khatri, S.M.; Babu, D.R.R.: A novel scheme for term weighting in text categorization: positive impact factor. Paper Presented at the 2013 IEEE International Conference on Systems, Man, and Cybernetics (2013)
21. Badawi, D.; Altınçay, H.: A novel framework for termset selection and weighting in binary text classification. Eng. Appl. Artif. Intell. **35**, 38–53 (2014). https://doi.org/10.1016/j.engappai.2014.06.012
22. Ke, W.: Information-theoretic term weighting schemes for document clustering and classification. Int. J. Digit. Libr. **16**(2), 145–159 (2015). https://doi.org/10.1007/s00799-014-0121-3
23. Deng, Z.-H.; Luo, K.-H.; Yu, H.-L.: A study of supervised term weighting scheme for sentiment analysis. Expert Syst. Appl. **41**(7), 3506–3513 (2014). https://doi.org/10.1016/j.eswa.2013.10.056
24. Abdel Fattah, M.: New term weighting schemes with combination of multiple classifiers for sentiment analysis. Neurocomputing **167**, 434–442 (2015). https://doi.org/10.1016/j.neucom.2015.04.051
25. Escalante, H.J.; García-Limón, M.A.; Morales-Reyes, A.; Graff, M.; Montes-y-Gómez, M.; Morales, E.F.; Martínez-Carranza, J.: Term-weighting learning via genetic programming for text classification. Knowl. Based Syst. **83**, 176–189 (2015). https://doi.org/10.1016/j.knosys.2015.03.025
26. Ko, Y.: A new term-weighting scheme for text classification using the odds of positive and negative class probabilities. J. Assoc. Inf. Sci. Technol. **66**(12), 2553–2565 (2015). https://doi.org/10.1002/asi.23338
27. Chen, K.; Zhang, Z.; Long, J.; Zhang, H.: Turning from TF-IDF to TF-IGM for term weighting in text classification. Expert Syst. Appl. **66**, 245–260 (2016). https://doi.org/10.1016/j.eswa.2016.09.009
28. Haddoud, M.; Mokhtari, A.; Lecroq, T.; Abdeddaïm, S.: Combining supervised term-weighting metrics for SVM text classification with extended term representation. Knowl. Inf. Syst. **49**(3), 909–931 (2016). https://doi.org/10.1007/s10115-016-0924-1
29. Kim, H.K.; Kim, M.: Model-induced term-weighting schemes for text classification. Appl. Intell. **45**(1), 30–43 (2016)
30. Sabbah, T.; Selamat, A.; Selamat, M.H.; Al-Anzi, F.S.; Viedma, E.H.; Krejcar, O.; Fujita, H.: Modified frequency-based term weighting schemes for text classification. Appl. Soft Comput. **58**, 193–206 (2017)
31. Badawi, D.; Altınçay, H.: Termset weighting by adapting term weighting schemes to utilize cardinality statistics for binary text categorization. Appl. Intell. (2017). https://doi.org/10.1007/s10489-017-0911-6
32. Wu, H.; Gu, X.; Gu, Y.: Balancing between over-weighting and under-weighting in supervised term weighting. Inf. Process. Manag. **53**(2), 547–557 (2017). https://doi.org/10.1016/j.ipm.2016.10.003
33. Alsmadi, I.; Hoon, G.K.: Term weighting scheme for short-text classification: twitter corpuses. Neural Comput. Appl. (2018). https://doi.org/10.1007/s00521-017-3298-8
34. Rao, Y.; Li, Q.; Wu, Q.; Xie, H.; Wang, F.L.; Wang, T.: A multi-relational term scheme for first story detection. Neurocomputing **254**, 42–52 (2017)
35. Feng, G.; Li, S.; Sun, T.; Zhang, B.: A probabilistic model derived term weighting scheme for text classification. Pattern Recognit. Lett. **110**, 23–29 (2018)
36. Matsuo, R.; Ho, T.B.: Semantic term weighting for clinical texts. Expert Syst. Appl. **114**, 543–551 (2018)
37. Li, X.; Zhang, A.; Li, C.; Ouyang, J.; Cai, Y.: Exploring coherent topics by topic modeling with term weighting. Inf. Process. Manag. **54**(6), 1345–1358 (2018)

38. Santhanakumar, M.; Columbus, C.C.; Jayapriya, K.: Multi term based co-term frequency method for term weighting in information retrieval. Int. J. Bus. Inf. Syst. **28**(1), 79–94 (2018)
39. Pak, A.; Paroubek, P.; Fraisse, A.; Francopoulo, G.: Normalization of term weighting scheme for sentiment analysis. In: Language and Technology Conference, pp. 116–128. Springer (2011)
40. Erenel, Z.; Altınçay, H.: Nonlinear transformation of term frequencies for term weighting in text categorization. Eng. Appl. Artif. Intell. **25**(7), 1505–1514 (2012). https://doi.org/10.1016/j.engappai.2012.06.013
41. Xuan, N.P.; Le Quang, H.: A new improved term weighting scheme for text categorization. In: Knowledge and Systems Engineering. Advances in Intelligent Systems and Computing, pp. 261–270. (2014). https://doi.org/10.1007/978-3-319-02741-8_23
42. Nguyen, T.T.; Chang, K.; Hui, S.C.: Supervised term weighting centroid-based classifiers for text categorization. Knowl. Inf. Syst. **35**(1), 61–85 (2013)
43. Lan, M.; Tan, C.L.; Su, J.; Lu, Y.: Supervised and traditional term weighting methods for automatic text categorization. IEEE Trans. Pattern Anal. Mach. Intell. **31**(4), 721–735 (2009). https://doi.org/10.1109/TPAMI.2008.110
44. Rocchio JJ (1971) Relevance feedback in information retrieval. In: The smart retrieval system-experiments in automatic document processing, pp 313–323
45. Chang, C.-C.; Lin, C.-J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) **2**(3), 27 (2011)
46. Porter, M.F.: An algorithm for suffix stripping. Program **14**(3), 130–137 (1980)
47. Asuncion, A.; Newman, D.J.: UCI Machine Learning Repository. http://www.ics.uci.edu/~mlearn/MLRepository.html. Accessed Jan 2013 (2007)